Article from

Risk Management

December 2016
Issue 37

Article from

Risk Management

December 2016
Issue 37

Estimating Probability of a Cybersecurity Breach

By Meghan Anthony, Maria Ishmael, Erik Santa, Arkady Shemyakin, Gary Stanull and Natalie Vandeweghe

Editor's Note

Cyber risk management has been integrated into companies' day-to-day operations. However, the evolving threats and fragmented data on cyber risk present a challenge for companies to understand and quantify a cybersecurity breach.

In this issue, we are pleased to share with readers a research paper from Professor Shemyakin and his team from University of St. Thomas on Estimating Probability of a Cybersecurity Breach. This article discusses how to estimate probability of a breach for a specific database application. In a simple example, the probability of a breach for a database with 100,000 records can be estimated by the probability of a database breach and a BF factor. The BF factor is derived from a predictive model as discussed below. This estimate would provide decision-makers information about the probability of a breach for a specific application, so to identify the most vulnerable applications, and make it possible to assign "risk ratings" on applications.

INTRODUCTION

Information technology is the engine that drives the U.S. economy, giving it a competitive advantage in global markets by providing better services and facilitating greater productivity. This great value means that information systems are subject to a variety of threats, from malicious hackers to an employee simply losing a flash drive. Unfortunately, the threat landscape is constantly changing. To determine the risk these threats pose we need to evaluate the likelihood of their success in exploiting known and unknown vulnerabilities. This involves an accurate assessment of both impact and probability of breaches.

The purpose of this paper is to define a predictive model, based on known system attributes, for assessing risk associated with information systems. The goal is to provide decision-makers with the best possible information about the probability of a security breach so they can make informed decisions on how to best address the risk.

Most statistical papers dealing with cybersecurity are dedicated to analysis of the breach data and development of distribution models for frequency of the breaches and severity of the associated losses. We will deal with a different problem: how to estimate probability of a breach for a specific database application. This estimate would make it possible to assign "risk ratings," allowing decision-makers to identify the most vulnerable applications based on some of their observable characteristics. However, obtaining such estimates will require us to build distribution models of these characteristics not only for "broken" applications, but also for "unbroken" ones.

To illustrate this point, let us consider an attribute A of an application (such as size, type of data, or the industry). Using the Bayes formula, we may assess the probability of an application with this attribute to be broken:

$$P(B/A)=\frac{P(B)P(A/B)}{P(B)P(A/B)+P(U)P(A/U)}, \quad (1)$$

where $P(B)$ is overall prior probability of an application to be broken, $P(U)=1-P(B)$; while $P(A/B)$ and $P(A/U)$ are the probabilities to observe attribute A among broken and unbroken applications respectively. The latter two can be estimated from the historical data when such data are available. This estimation requires certain knowledge regarding the population of unbroken applications, which is not typically used in analysis of the breach data.

PROBABILITY OF BREACH GIVEN APPLICATION SIZE

For an illustration let us concentrate on health-related data and single out such application attribute as record size measured as the total number of records. We will be considering the following grouping of application record sizes:

S_1 : Below 10,000

S_2 : 10,000–30,000

S_3 : 30,000–100,000

S_4 : 100,000–1 million

S_5 : Above 1 million

Suppose that independently of the application size, the probability of that to be breached (or broken) is estimated as $P(B)$, the prior probability of a breach. Then the posterior probability of a breach conditional on the size of the class S_k, where $k=1,2,3,4,5$, can be evaluated as

$$P(B/S_k)=\frac{P(B)P(S_k/B)}{P(B)P(S_k/B)+(1-P(B))P(S_k/U)}, \quad (2)$$

where $P(S_k / B)$ and $P(S_k / U)$ will be estimated from two different parametric distribution models for the random size X_B of a broken application $F_B(x) = P(X_B \le x)$ and the size X_U of an unbroken application $F_U(x) = P(X_U \le x)$:

$$P(S_k / B) = \int_{S_k} x dF_B(x),\ P(S_k / U) = \int_{S_k} x dF_U(x). \quad (3)$$

DISTRIBUTION MODELS FOR APPLICATION SIZE

For the data on breaches, we used 1,572 data points recorded in the HHS database [2]. For the data on the unbroken applications we used a sample size of 81 obtained from industry experience of application analysis from 2014 to 2015.

Application size is measured as the total number of records R log-transformed and shifted with natural thresholds 500 (minimum record size for broken data) and 10 (minimum record size for unbroken data) so that $X_B = \ln(R) - \ln(499)$ and $X_U = \ln(R) - \ln(9)$

For variables X_B and X_U three two-parameter distribution models are considered: normal (Gaussian), suggested in Edwards, Hofmeyr and Forrest [1]; gamma; and Weibull. Maximum likelihood estimates (MLE) were obtained for all three models. In Figure 1 and Figure 2 the boundary of the shaded area corresponds to empirical CDF, the best normal fits are depicted by dashed lines, fitted gamma in red and fitted Weibull in green.

Figure 1. Distribution $F_B(x) = P(X_B \le x)$

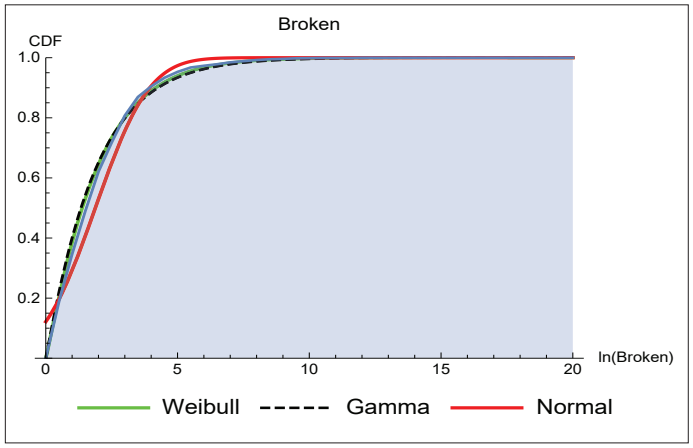
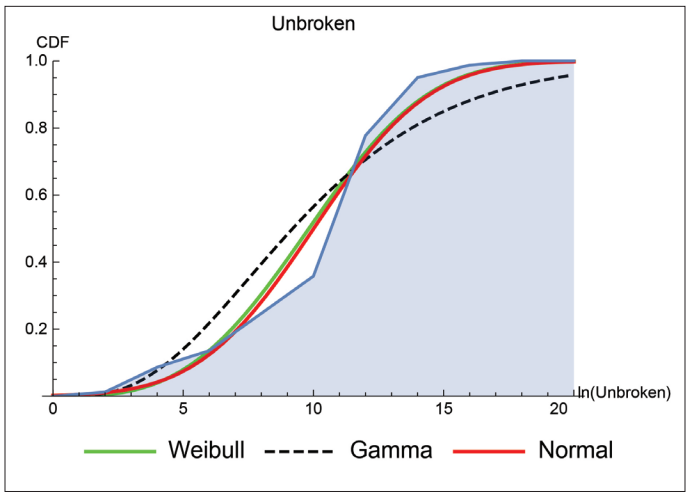


Figure 2. Distribution $F_U(x) = P(X_U \le x)$



Model selection using Akaike information criterion (AIC) brings about the results in Table 1.

Table 1. AIC Values (the lower the better)

Model	Normal	Gamma	Weibull
X_B	5964.4	5129.9	**5117.2**
X_U	**435.5**	479.3	447.6

Based on AIC values from Table 1 and graphs of $F_B(x)$ in Figure 1 and $F_U(x)$ in Figure 2, we recommend the choice of Weibull distribution for both cases: $F_M(x) = 1 - e^{-(x/\theta_M)^{\tau_M}}$ with scale θ_M and shape τ_M estimated separately for models $M = B, U$ via MLE:

$$\tau_B = 1.089,\ \theta_B = 1.937;\ \tau_U = 3.156,\ \theta_U = 11.036.$$

BAYES FACTORS

Based on the results from the previous section, we can estimate $P(S_k / B)$ and $P(S_k / U)$. Obtaining posterior probabilities of a breach directly from Eq. (2) would require an additional specification of the prior probability $P(B)$, generally unknown. Therefore, we will use *Bayes factors* defined for each $k = 1,\ldots,5$ as

$$BF_k = \frac{P(S_k / B)}{P(S_k / U)}$$

It follows from Eq. (2) that for sufficiently small $P(B)$ the Bayes factors approximate the ratios of probabilities and thus represent appropriate "adjustments" to prior probabilities of

$$BF_k \approx \frac{P(B / S_k)}{P(B)}$$

breach, which take into account the application size. The results are summarized in Table 2.

Table 2. Conditional Probabilities and Bayes Factors.

S_k	$P(S_k / B)$	$P(S_k / U)$	BF_k
≤ 10,000	0.800	0.213	3.756
10,000 ≤ 30,000	0.096	0.102	0.941
30,000 ≤ 100,000	0.054	0.128	0.422
100000 ≤ 1,000,000	0.038	0.248	0.153
≥ 1,000,000	0.011	0.308	0.036

These results can be interpreted as follows: When we start with a small prior probability of a breach $P(B)$, additional knowledge that the application size belongs to a certain class S_k, makes it possible to estimate the posterior probability as $P(B / S_k) \approx P(B) \times BF_k$. According to the first and the last rows of Table 2, small size of an application makes it more likely to be broken, while very large size makes it much less vulnerable. This may be characteristic for the breaches of healthcare applications being often caused by thefts of laptops and storage devices, which are unlikely to contain large size applications.

CONCLUSIONS:

- Weibull distribution provides the best overall fit for B and U
  - Data points in the tail of the dataset (large record sizes) do not deviate from the best-fit trend-line as was noticed in Edwards, Hofmeyr and Forrest [1] for Gaussian model.
- Bayes factors can be used to evaluate posteriors for small prior probabilities $P(B)$.
- Probability of data breach can be effectively adjusted based on the record size.

Next Steps:

- Perform similar analysis and calculate Bayes factors based on the record size for different breach types (e.g., theft of a laptop versus a large-scale intentional hacking).
- Identify entity attributes such as revenue that correlate with the record size.
  - Use these attributes as proxies to model distribution of record sizes for "unbroken" applications.

- Analyze the Privacy Rights Clearinghouse dataset [6] to expand the scope of analysis to the industries beyond healthcare.
- Estimate risk to information systems based on the record size.

REFERENCES

[1] Edwards, B., Hofmeyr, S., Forrest, S. (2015). Hype and Heavy Tails: A Closer Look at Data Breaches, *http://www.econinfosec.org/archive/weis2015/papers/WEIS_2015_edwards.pdf* .

[2] United States Department of Health and Human Services (HHS), *www.hhs.gov* .

[2] Verizon Data Breach Incident Reports (DBIR), 2008 – 2016, *www.verizonenterprise.com/verizon-insights-lab/dbir/*.

[3] Symantec Healthcare Internet Security Threat Report (ISTR), 2016, *www.symantec.com* .

[4] Ponemon Institute 2015 Fifth Annual Benchmark Study on Privacy & Security of Healthcare Data, *www.ponemon.org*.

[5] HITRUST A Look Back: U.S. Healthcare Data Breach Trends 2012, *hitrustalliance.net* .

[6] Privacy Rights Clearinghouse Chronology of Data Breaches and Security Breaches 2005 – Present, *www.privacyrights.org/*. ■

Meghan E. Anthony, University of St. Thomas, *anth4118@stthomas.edu*.

Arkady Shemyakin, Ph.D., is professor and director of statistics program at University of St. Thomas. He can be reached at *a9shemyakin@stthomas.edu*.

Maria L. Ishmael, University of St. Thomas, *ishm3313@stthomas.edu.*

Gary Stanull, BS, MBA, MS, CISSP, CISM, is enterprise security architect at Optum Technology. He can be reached at *gary.stanull@optum.com*.

Erik W. Santa, University of St. Thomas, *sant5579@stthomas.edu*.

Natalie A. Vandeweghe, University of St. Thomas, *vand1269@stthomas.edu.*